INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C07H 21/04, C12N 15/00 | A1

(11) International Publication Number: WO 99/22025

(43) International Publication Date: 6 May 1999 (06.05.99)

(21) International Application Number: PCT/US98/22519

(22) International Filing Date: 23 October 1998 (23.10.98)

(30) Priority Data:

60/063,103	24 October 1997 (24.10.97)	US
Not furnished	22 October 1998 (22.10.98)	US

(71) Applicant: TRUSTEES OF BOSTON UNIVERSITY [US/US]; 147 Bay State Road, Boston, MA 02215 (US).

(72) Inventors: CANTOR, Charles, R.; 11 Bay State Road #6, Boston, MA 02115 (US). SIDDIQI, Fouad, A.; Apartment #3, 113 Bay State Road, Boston, MA 02115 (US).

(74) Agents: CARROLL, Peter, G. et al.; Medlen & Carroll, LLP, Suite 2200, 220 Montgomery Street, San Francisco, CA 94104 (US).

(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published

With international search report.

(54) Title: INFERENCE SEQUENCING BY HYBRIDIZATION

(57) Abstract

The inventors present a new strategy for *de novo* sequencing using oligonucleotide probe arrays which we term Inference Sequencing by Hybridization (ISBH). The ISBH method characterises a target DNA with several small degenerate probe arrays and reconstructs the base sequence of the target via computer algorithm at 100 % accuracy. ISBH returns a target sequence as multiple non-overlapping fragments several hundred to several thousand bases in length. Target DNAs up to 100 kilobases long can potentially be sequenced in a single experiment. A hypothetical ISBH probe array and sequence reconstruction algorithm were designed and tested by computer simulation on 76 DNA sequences obtained from GenBank comprising a total of 2.45 million base pairs. ISBH performs optimally on target DNAs in the range of 15–45 kilobases that have low levels of repeated sequence. We believe that the ISBH approach has great potential to increase the speed of DNA sequencing.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

INFERENCE SEQUENCING BY HYBRIDIZATION

This application for patent under 35 U.S.C. § 111(a) claims priority to Provisional Application Serial No. 60/063,103, filed October 24, 1997 under 35 U.S.C. § 111(b). This invention was made with Government Support under Contract Number DGE-9452651 awarded by the National Science Foundation. The Government has certain rights in the invention.

BACKGROUND OF THE INVENTION

Procedures involving use of Sequencing by hybridization (SBH) are known to those skilled in the art, and have recently been demonstrated to be useful as a powerful alternative to electrophoretic methods for diagnostic DNA analysis [M. Chee *et al.*, Accessing Genetic Information With High Density DNA Arrays, *Science* 274, 610 (1996); J. Hacia *et al.*, Detection Of Heterozygous Mutations In BRCA1 Using High-Density Oligonucleotide Arrays And Two-Color Fluorescence Analysis, *Nature Genetics* 14, 441 (1996)]. Diagnostic SBH employs hybridization of a target DNA sequence to a tiled array of several thousand short oligonucleotide probes of known sequence [W. Bains *et al.*, A Novel Method For Nucleic Acid Sequence Determination, *J Theor Biol*, 135, 303 (1988); E. Southern *et al.*, Hybridization With Oligonucleotide Arrays, *Genomics*, 13, 1008 (1992)]. The pattern of hybridization, detected by fluorescence microscopy, indicates which oligonucleotides in the probe array are present in the target DNA. When this information is compared against a reference target sequence, the entire sequence of the target DNA can be reconstructed at high accuracy.

SBH, while offering great advantages in terms of throughput for diagnostic sequence analysis, suffers from the drawback that a different probe array must be tailored for each target DNA analyzed [M. Chee *et al.*, Accessing Genetic Information With High Density DNA Arrays, *Science*, 274, 610 (1996)]. For *de novo* sequencing, current SBH methods are not competitive with electrophoretic sequencing techniques

that yield 600-1000 base pair read lengths per experiment [P. Pevzner *et al.*, Improved Chips For Sequencing By Hybridization, *J Biomolecular Struct Dyn*, 9, 399 (1991)]. Even under perfect experimental conditions, existing SBH designs cannot reconstruct a unique target sequence from hybridization data alone [P. Pevzner *et al.*, Towards DNA Sequencing Chips, In *19th Int Conf Mathematical Foundations of Computer Science*, Lecture Notes Iin Computer Science, Springer-Verlag, Berlin, Vol 841, pp. 143-158 (1994); P. Pevzner, Rearrangements Of DNA Sequences And SBH, *Computers Chem*, 18, 221 (1994)]. Without a reference sequence for comparison, *de novo* SBH is fundamentally limited because it acquires base sequence information at the cost of positional information. One knows exactly *which* subsequences (probe sequences) are present in the target DNA but not *where* they are located.

Subsequences must be arranged by examining how they overlap with one another. For example, octa-nucleotides are assembled into longer sequences by finding corresponding seven-base overlaps. The accuracy of reassembly is limited because any particular subsequence can occur more than once in the target DNA, leading to an ambiguity in the final reconstructed sequence [P. Pevzner *et al.*, Towards DNA Sequencing Chips, In *19th Int Conf Mathematical Foundations of Computer Science*, Lecture Notes Iin Computer Science, Springer-Verlag, Berlin, Vol 841, pp. 143-158 (1994); P. Pevzner, Rearrangements Of DNA Sequences And SBH, *Computers Chem*, 18, 221 (1994)]. *De novo* SBH designs typically use the complete set of all possible oligonucleotide probes of a given length [W. Bains *et al.*, A Novel Method For Nucleic Acid Sequence Determination, *J Theor Biol*, 135, 303 (1988); N. Broude *et al.*, Enhanced DNA Sequencing By Hybridization, *Proc Natl Acad Sci USA*, 91, 3072 (1994); R. Drmanac *et al.*, DNA Sequence Determination by Hybridization: A Strategy For Efficient Large-Scale Sequencing, *Science*, 260, 1649 (1993); R. Drmanac *et al.*, Sequencing Of Megabase Plus DNA By Hybridization: Theory Of The Method, *Genomics*, 4, 114 (1989)]; the use of longer probes increases reconstruction accuracy but requires the use of very large arrays ($> 10^9$ probes), since the number of required probes increases exponentially with probe length. Even assuming perfect hybridization,

an SBH array containing all ~10^6 possible 10-mers would reliably be able to sequence only about 600 bp of target DNA in a single experiment. As longer target DNA sequences are attempted, the reconstruction accuracy drops precipitously. For a detailed discussion of SBH reassembly algorithms and their limitations see references [P. Pevzner *et al.*, Towards DNA Sequencing Chips, In *19th Int Conf Mathematical Foundations of Computer Science*, Lecture Notes Iin Computer Science, Springer-Verlag, Berlin, Vol 841, pp. 143-158 (1994); P. Pevzner *et al.*, Improved Chips For Sequencing By Hybridization, *J Biomolecular Struct Dyn*, 9, 399 (1991); P. Pevzner, Rearrangements Of DNA Sequences And SBH, *Computers Chem*, 18, 221 (1994)].

All *de novo* SBH strategies proposed thus far are direct methods, that is, they directly probe the target DNA with oligonucleotides whose sequences are then assembled into longer fragments. Since it is currently not feasible to manufacture a probe array containing more than approximately 10^6 tiled oligonucleotide probes, a direct *de novo* SBH approach cannot outperform electrophoretic sequencing in terms of read length and reaaccuracy.

IN THE FIGURES

The present invention will become better understood with reference to the following description, appended claims, and accompanying figures where:

Figure 1. Basic concepts underlying the design of an Inference Sequencing by Hybridization (ISBH) probe array. A target, in this case a four-digit phone number, is characterized by a degenerate probe array with a four-fold redundancy using only 64 probes. The probe array does not detect any digit directly, but the information gathered is sufficient to unambiguously infer the identity of the number. A probe array based on conventional SBH designs capable of acquiring the same information would require 10,000 probes, one for each phone number.

Figure 2. General scheme for ISBH. A long single-stranded target DNA is sheared into short oligonucleotides and hybridized to an ISBH probe array. The pattern of hybridization is used to create a set of degenerate 16-mers that characterize

the target DNA. Information from this degenerate set is used by an inference algorithm to produce a set of explicit 16-mers. The set of explicit 16-mers produced by the inference algorithm contains all 16-mers actually present in the target DNA sequence as well as "false positive" 16-mers that are not in the target. A data reduction algorithm is then used to eliminate the false positives from the set of explicit 16-mers. The explicit 16-mers that remain after data reduction are then reassembled at high accuracy into contiguous sequence.

Figure 3. Design of the ISBH probe array of degenerate 16-mers used in this study. The array consists of 25 different probe groups. Each group pattern represents $2^{16} = 65,536$ degenerate 16-mers, for a total of 1,638,400 probes. Each probe in the array represents 65,536 explicit 16-mers. Under ideal conditions, a single target 16-mer will hybridize to exactly one probe in each probe group.

Figure 4. Example of how false positives are generated by inference. Tetranucleotides from a target DNA are characterized by an ISBH probe array of degenerate 4-mers with two-fold redundancy (R = A or G; Y = C or T; W = A or T; S = C or G). The inference algorithm generates all valid combinations of data from the probe array to produce a set of nine inferred tetranucleotides, six of which are false positives. In general, the number of false positives generated by inference decreases with the number of probes used in the ISBH array. If the ISBH array used in this example had 16 additional probes, then no false positives would have been generated.

Figure 5. Data reduction after inference for the ISBH array of degenerate 16-mers used in this study. Seventy-six target DNA sequences downloaded from GenBank comprising a total of 2.45 million bases were tested by computer simulation. The number of 16-mers generated by inference increases as a power law function of the number of different 16-mers in the target DNA (filled circles). Data reduction reliably eliminated all but a handful of the false positives for all target lengths investigated (open triangles), even when false positives comprised more than 99% of the inferred 16-mer set.

Figure 6. Absolute performance of the ISBH sequence reassembly algorithm. The reassembly algorithm returns a target DNA as several non-overlapping fragments

in the range of several hundred to several thousand bases in length. The largest fragment reconstructed in this study was 28 kilobases, and fragments longer than 10kb were commonly observed. Reconstructed fragments always show 100% identity to some region of the target DNA.

Figure 7. Total target coverage in a simulated ISBH experiment. ISBH typically covers more than 95% of a target DNA in a single hybridization experiment in fragments that are longer than 500 bases.

Figure 8. Summary of ISBH simulation results. All lengths are in bases. Locus and definition for each sequence are shown exactly as they appear in GenBank. The number different 16-mers in a target is defined as the number of 16-mers which have different base sequences. The fraction of target 16-mers that are repeated is given by: 1 - [(number of different 16-mers in target)/(target length - 15)], which is a quantitative measure of repetitiveness of the sequence. The fraction of the inferred set that are false positive is given by: 1 - [(number of different 16-mers in target)/(number of 16-mers in inferred set)].

SUMMARY OF THE INVENTION

The present invention is directed to a method that satisfies the above mentioned problems by introducing a new SBH implementation for *de novo* sequencing, which we term Inference Sequencing by Hybridization (ISBH). The basic concepts underlying ISBH are illustrated in Figure 1 with an example of how to determine the last four digits of a phone number without detecting any digit directly. A conventional SBH strategy would require $10^4 = 10{,}000$ probes (one for each phone number), whereas the ISBH approach gathers the same information using only $4 \times 2^4 = 64$ degenerate probes. The digit groupings in Figure 2 are analogous to familiar base groupings such as purine/pyrimidine (R/Y), amino/keto (M/K), and weak/strong (W/S).

ISBH is an indirect strategy that uses several small arrays (65,536 probes each) to closely approximate the information that would be gathered from a single SBH array containing ~4.3 billion probes (Figure 2). Our strategy relies on degenerate probe arrays that are similar to binary SBH arrays proposed by Pevzner *et. al.* [P. Pevzner *et*

al., Towards DNA Sequencing Chips, In *19th Int Conf Mathematical Foundations of Computer Science*, Lecture Notes Iin Computer Science, Springer-Verlag, Berlin, Vol 841, pp. 143-158 (1994)]. Unlike conventional SBH, whose accuracy drops with increasing target length, we find that ISBH always reconstructs targets at 100% accuracy. ISBH returns a target sequence as non-overlapping fragments that range from several hundred to several thousand bases in length. The number of fragments increases with the length and repetitiveness of the target DNA, but in principle, any length of target can be sequenced in a single experiment. We have designed and empirically tested by computer simulation an ISBH array and reconstruction algorithm for use in *de novo* sequencing of targets up to 100 kilobases in length.

METHODS

The inventors have simulated the following laboratory experiment: 1) a single stranded target DNA of unknown sequence is sheared into overlapping oligomers 16 bases long. 2) These oligomers are hybridized to an ISBH degenerate probe array and this pattern of hybridization is detected. 3) the hybridization data is reconstructed by computer algorithm into contiguous sequence.

An ISBH probe array containing 1.64 million different degenerate 16-mers was designed. This array is 25-fold redundant - it consists of 25 groups of 2^{16}=65,536 (64K) different probe oligomers. Figure 3 shows the identity of each degenerate probe group in the array. Each group of 64K degenerate probes is capable of hybridizing to all possible explicit 16-mers. Thus, a single explicit 16-mer will, under ideal conditions, hybridize to exactly 25 different degenerate probes in the ISBH array.

Simulations of ISBH experiments and subsequent data analysis were performed on a Silicon Graphics Origin2000 supercomputer (Boston University Center for Computational Science) with code written in C as follows: 1) Each target DNA sequence was retrieved from GenBank (http:// *www2.ncbi.nlm.nih.gov/genbank*) and broken up into all possible component 16-mers. This set of 16-mers was then shuffled to destroy any positional information. 2) Each 16-mer from the target DNA was compared against all 1.64 million degenerate probes in the ISBH array, simulating

ideal hybridization. If a 16-mer from the target was found to hybridize with one of the degenerate probes, then the hybridization was considered to be a signal arising from that particular probe in the ISBH array. 3) The 16-mers from the target DNA were then discarded, to account for the fact that in a real experiment, the target DNA is of unknown sequence. 4) The signals from the ISBH array were collected to produce a set of degenerate 16-mers, which is the non-repeated set of degenerate 16-mers present in the target DNA. We emphasize that the only data retained is the signal pattern from the ISBH array, as would be detected in an actual ISBH experiment. The degenerate set contains no information about the explicit base sequence of any 16-mer present or its position in the target DNA sequence.

Sequence Reconstruction

1) *Inference.* A set of explicit 16-mers is inferred from the set of degenerate 16-mers detected by the ISBH array. The inference is accomplished by testing every possible explicit 16-mer against the degenerate set. Any particular explicit 16-mer is included in the inferred set only if exactly 25 corresponding degenerate 16-mers are present in the data from the ISBH array. Under conditions of ideal hybridization, the inferred 16-mer set is always a superset of the set of 16-mers present in the target DNA sequence. The inferred set usually contains false positive 16-mers, ones which are not actually present in the original target DNA sequence. The number of these false positives increases with the length and repetitiveness of the target DNA (Figure 4).

2) *Primary Data Reduction.* Since the inferred set of explicit 16-mers contains an unknown number of false positives, a data reduction step is required to eliminate them. Every 16-mer in the inferred set is examined to determine if it overlaps by six bases at least one other 16-mer in the inferred set on both its 3' and 5' ends. If both overlaps are not found the 16-mer is discarded. This procedure is repeated iteratively on the resultant set of 16-mers, each time examining an overlap one base longer than was used for the previous iteration, until an overlap of fifteen bases is reached. The

set that remains is then iterated using fifteen-base overlaps until four or fewer 16-mers are discarded at each iteration.

3) *Secondary Data Reduction.* All possible reconstruction ambiguities are eliminated by comparing the 3'-fifteen bases of each 16-mer with the 3'-fifteen bases of every other 16-mer in the set from step 2 above. If two or more 16-mers are found to have identical fifteen base 3'-ends, then they are all discarded from the data set. A similar procedure is used to compare the 5'-fifteen bases of the 16-mers and eliminate any duplication.

4) *Sequence Reassembly.* The 16-mers remaining in the inferred set are then assembled into longer sequences. This is done by comparing the 3'-fifteen bases of each 16-mer to the 5'-fifteen bases of every other 16-mer in the data set. If a match is found, then the two 16-mers are combined into a single 17-mer. The two terminal 15-mers on either side of this newly formed 17-mer are compared for overlap with the remaining 16-mers in the data set and the process is repeated until no more overlaps are found. To insure reconstruction accuracy, only fragments at least 100 bases in length were considered to be part of the target sequence.

RESULTS

Inference Algorithm

ISBH simulations were performed on 76 target sequences obtained from GenBank comprising a total of 2.45 million bases, ranging from 5 to 100 kilobases in length. The size of the inferred 16-mer pool increases as a power law function of the number of different 16-mers in the target DNA. Data reduction reliably eliminates all but a handful of the false positives for all target lengths investigated, even when false positives accounted for more than 99% of the inferred 16-mer pool (Figure 5) . The set of inferred 16-mers remaining after data reduction closely approximates the information that would be gathered from an SBH array containing all explicit 16-mers (~4.3 billion probes).

Sequence Reconstruction

The ISBH reconstruction algorithm typically returns a target DNA as several non-overlapping fragments that are in the range of several hundred to several thousand bases in length. Reconstructed fragments always show 100% identity (by BLASTN) to some region of unknown position in the target DNA sequence. The largest single fragment reconstructed was 28 kb, and fragments longer than 10kb were commonly observed (Figure 6). In most cases, more than 95% of the target DNA was recovered in a simulated ISBH experiment (Figure 7). ISBH performs optimally on target DNAs having no repeated 16-mers, generally returning a handful of long (3-15 kb) fragments. Even on sequences with many repeated 16-mers, ISBH returns dozens of fragments shorter than 5 kb, which is five to ten times the performance of electrophoretic sequencing methods. Target DNAs longer than 50 kb tend to produce large numbers of false positives in the inference step, a few of which remain after data reduction. False positives introduce ambiguities during reassembly, leading to lower average lengths for reconstructed fragments. A comprehensive list of each sequence analyzed as well as a summary of the ISBH simulation results are shown in Figure 8.

DETAILED DESCRIPTION OF THE INVENTION

While this invention is satisfied by embodiments in many different forms, there will herein be described preferred embodiments of the invention, with the understanding that the present disclosure is to be considered exemplary of the principles of the invention and is not intended to limit the invention to the embodiments illustrated and described. The scope of the invention will be measured by the appended claims and their equivalents.

A benefit of the present invention, in contrast to conventional SBH, is that the inventor's ISBH method has the potential to sequence very long targets at high accuracy, using an oligonucleotide array of moderate size. The hypothetical ISBH array studied here could easily sequence 15-45kb of DNA in a single experiment. The ISBH method requires no electrophoresis, no information about the target DNA, and could be used for diagnostic as well as *de novo* applications. In a single experiment,

ISBH could generate more sequence data than two dozen Sanger sequencing reactions after shotgun subcloning of a target DNA. In the best cases, each fragment reconstructed by the ISBH method can outperform electrophoretic methods by 28-fold. In the worst cases, ISBH performance is equivalent to electrophoretic methods. ISBH reconstruction of a single target DNA generally required less than 10 minutes of supercomputer time to complete. The computational complexity of the inference step is of order N^2, while the data reduction and reassembly steps are of order N log(N). Sequence reconstruction using a highly streamlined ISBH algorithm running on a typical desktop computer could be completed in a few hours.

For DNA of random sequence, a given 16-mer should appear once every $4^{16} \approx 4.3 \times 10^9$ bases, a 15-mer once in $4^{15} \approx 10^9$ bases, and a 14-mer once in $4^{14} \approx 2.7 \times 10^8$ bases. We note however, that DNAs from a wide variety of organisms in the range of 10-100 kb typically have hundreds or thousands of repeated 16-mers. As shorter subsequences are examined, the number of repeated subsequences increases dramatically. For example, the 48.5 kb genome of bacteriophage lambda, which has no repeated 16-mers, has a single repeated 15-mer, and ten distinct 14-mers appearing more than once. This would suggest that any form of *de novo* SBH using oligonucleotide probes shorter than 16 bases will perform poorly on target DNAs longer than a few kilobases. ISBH appears to perform optimally both in terms of absolute read lengths and relative target coverage on DNAs in the range of 25kb that have small numbers of repeated 16-mers. ISBH is a scalable technique - the number of false positives generated by inference increases as the number of probes used in the ISBH array decreases. An ISBH array smaller than the one examined here (*e.g.*, 12 probe groups using $12 \times 2^{16} = 7.86 \times 10^5$ probes) would still sequence with 100% accuracy, but would return a target as shorter fragments.

While ISBH under ideal conditions would appear to provide an enormous gain for *de novo* sequencing over conventional SBH and electrophoretic methods, several daunting technical obstacles remain. Each degenerate probe is actually a mixture of many individual probes that are bound to the same area in an SBH array. The binding capacity of such a degenerate probe is greatly reduced in comparison to a pure

individual probe - for the hypothetical ISBH array in this study, the complexity of each degenerate probe is 65,536. Such a high probe complexity may mean that accurate physical hybridization cannot be achieved with a high signal to noise ratio. Possible solutions to this problem include the use of base analogs to decrease probe complexity or the addition of an enzymatic step (*e.g.*, ligation) to augment the accuracy of simple physical hybridization.

Noise contamination of the data set, particularly in terms of false negatives, must be studied in greater detail. False positives are easily dealt with in the data reduction step, but false negatives (target 16-mers that never appear at all in the inferred data set) will have the effect of lowering the mean fragment length during reconstruction. Aberrant hybridization also increases the complexity of data processing needed for reliable sequence reconstruction; the upper bound for robust sequence reconstruction from an actual ISBH implementation is likely to be somewhat lower than the ideal situation presented here.

An ISBH sequencing approach would be very effective for rapid analysis of viral and bacterial genomes which are essentially non-repeating. Sequencing of double-stranded target DNAs by ISBH is also possible, as is sequencing of a mixture of targets. For double-stranded DNA, ISBH performance is equivalent to the case of a single-stranded DNA twice as long. If a double stranded target is cleaved by a restriction endonuclease before hybridization, ISBH will return the sequence of each restriction fragment. If this experiment is repeated using a restriction endonuclease with a different recognition site, then the fragments can be aligned relative to one another using standard contig reassembly algorithms. The potential of the ISBH strategy is so strong that we are now investigating strategies to implement it in practice.

Another embodiment of the ISBH Probe Array, Experimental, and Data Analysis Algorithm Design consists of the following:

Probe Array Design. The proposed array consists of 768K oligonucleotide probes divided into three groups: 1) all 256K possible 9-base single-stranded

sequences, 2) all 256K possible 9-base 5'-overhanging partial duplexes, and 3) all 256K possible 9-base 3'-overhanging partial duplexes.

Target Preparation. The target DNA must be single stranded - it may be prepared (for example) by long PCR of a double stranded target using one primer that is biotinylated at its 5' end to facilitate purification from the other strand. The biotinylated strand may be captured on streptavidin-coated beads or column.

Experimental Design. Eight distinct oligonucleotides in two groups are required as follows. Group I: 5'-ANNNNNNN-3', 5'-CNNNNNNN-3', 5'-GNNNNNNN-3', 5'-TNNNNNNN-3'. Group II: 5'-P-NNNNNNNA-Sdd-3', 5'-P-NNNNNNNC-Sdd-3', 5'-P-NNNNNNNG-Sdd-3', and 5'-P-NNNNNNNT-Sdd-3'. P denotes a phosphate group, Sdd denotes a 3'-dideoxy base connected by a phosphorothioate linkage. Sixteen separate reactions are then performed: each oligonucleotide from group I is combined with an oligonucleotide from group II and then hybridized to the single-stranded target under conditions favoring accurate base-pairing. After hybridization has occurred, the oligonucleotides that are still remaining in solution (for example) must be removed by size-exclusion column chromatography. DNA ligase (and all necessary cofactors) are added to the reaction mixture. After the ligation reaction, exonuclease III is added to the reaction mixture to destroy any unligated oligonucleotides that are still hybridized to the target.

Under ideal conditions, the ligation products will be 16-mers of the form: 5'-ANNNNNNNNNNNNNNA-Sdd-3' (and all 15 other permutations of the end bases). The ligation products from each of the sixteen reactions are then hybridized separately to a probe array as described above.

Data Analysis. Each of the sixteen probe array hybridization experiments described above generates the following data: a set of 9-mers from probe group 1, a set of 9-mers from probe group 2, and a set of 9-mers from group 3. The following algorithm is used to expand the data: compare each 9-mer from group 2 to each 9-mers from group 3. If the 9-mer from group 2 has the same 3' two bases as the 5' two bases of the 9-mer in group 3 (*i.e.*, they have a two-base overlap), then combine the two 9-mers to form a single 16-mer (concatenate the 3' seven bases of the group 3

oligo to the 3' end of the group 2 oligo). This newly formed 16-mer is retained only if all eight of its 9-base subsequences are present in group 1. This analysis is performed for all probe array experiments and all retained 16-mers are collected into a single set. This set of 16-mers (which is the inferred set of explicit 16-mers from the target) is then subjected to the same data reduction and sequence reconstruction algorithms that we have previously described for ISBH.

Accordingly, this invention is not limited to the particular embodiments disclosed, but is intended to cover all modifications that are within the spirit and scope of the invention as defined by the appended claims.

CLAIMS

What is claimed is:

1. A method of testing a nucleic acid target, comprising:

a) fragmenting a single-stranded target DNA into single-stranded target DNA fragments;

b) testing said fragments so as to generate a signal for each fragment;

c) calculating a first set of N-mers from said signals, each of said N-mers having a sequence with 3' and 5' ends;

d) comparing a portion of the nucleic acid sequence of each of said N-mers of said first set with a portion of the nucleic acid sequence of every other N-mer in said first set for sequence overlap; and

e) eliminating each N-mer that is found not to display said overlap, so as to create a second set of N-mers.

2. The method of Claim 1, further comprising the steps:

f) comparing a portion of the nucleic acid sequence of each of said N-mers in said second set with a portion of the nucleic acid sequence of every other N-mer in said second set for sequence overlap, wherein the portion compared has a length in bases defined by N-1; and

g) identifying an instance where said portion of one N-mer from said second set is found to overlap with said portion of another N-mer from said second set, thereby identifying first and second overlapping N-mers.

3. The method of Claim 1, wherein non-target DNA fragments are added to said single-stranded target DNA fragments prior to step (b).

4. The method of Claim 2, wherein N is 16 and said portion compared in step (f) is fifteen bases in length.

5. The method of Claim 4, wherein said fifteen bases compared in step (f) are the 3'-fifteen bases of each 16-mer with the 5'-fifteen bases of every other 16-mer in said second set.

6. The method of Claim 5, further comprising the step (h) constructing a 17-mer from the combination of the sequences of said first and second overlapping N-mers of step (g).

7. A method of testing a nucleic acid target, comprising:

a) fragmenting a single-stranded target DNA into single-stranded fragments;

b) testing each of said fragments so as to generate a signal for each fragment;

c) calculating a first set of N-mers from said signals, each of said N-mers having a sequence with 3' and 5' ends;

d) comparing a portion of the nucleic acid sequence of each of said N-mers of said first set with a portion of the nucleic acid sequence of every other N-mer in said first set for sequence overlap;

e) eliminating each N-mer that is found not to display said overlap, so as to create a second set of N-mers;

f) comparing a portion of the nucleic acid sequence of each of said N-mers in said second set with a portion of the nucleic acid sequence of every other N-mer in said second set for sequence overlap, wherein the portion compared has a length in bases defined by N-1; and

g) identifying an instance where said portion of one N-mer from said second set is found to overlap with said portion of another N-mer from said second set, thereby identifying first and second overlapping N-mers.

8. The method of Claim 7, wherein non-target DNA fragments are added to said single-stranded target DNA fragments prior to step (b).

9. The method of Claim 8, wherein N is 16 and said portion compared in step (f) is fifteen bases in length.

10. The method of Claim 9, wherein said fifteen bases compared in step (f) are the 3'-fifteen bases of each 16-mer with the 5'-fifteen bases of every other 16-mer in said second set.

11. The method of Claim 10, further comprising the step (h) constructing a 17-mer from the combination of the sequences of said first and second overlapping N-mers of step (g).

12. A method of testing a nucleic acid target, comprising:

a) fragmenting a single-stranded target DNA into single-stranded fragments;

b) adding non-target DNA fragments to said single-stranded target DNA fragments to create a mixture;

c) testing each of said fragments so as to generate a signal for each fragment;

d) calculating a first set of N-mers from said signals, each of said N-mers having a sequence with 3' and 5' ends;

e) comparing a portion of the nucleic acid sequence of each of said N-mers of said first set with a portion of the nucleic acid sequence of every other N-mer in said first set for sequence overlap;

f) eliminating each N-mer that is found not to display said overlap, so as to create a second set of N-mers;

g) comparing a portion of the nucleic acid sequence of each of said N-mers in said second set with a portion of the nucleic acid sequence of every other N-mer in said second set for sequence overlap, wherein the portion compared has a length in bases defined by N-1; and

h) identifying an instance where said portion of one N-mer from said second set is found to overlap with said portion of another N-mer from said second set, thereby identifying first and second overlapping N-mers.

13. The method of Claim 12, wherein N is 16 and said portion compared in step (g) is fifteen bases in length.

14. The method of Claim 13, wherein said fifteen bases compared in step (g) are the 3'-fifteen bases of each 16-mer with the 5'-fifteen bases of every other 16-mer in said second set.

15. The method of Claim 14, further comprising the step (i) constructing a 17-mer from the combination of the sequences of said first and second overlapping N-mers of step (h).

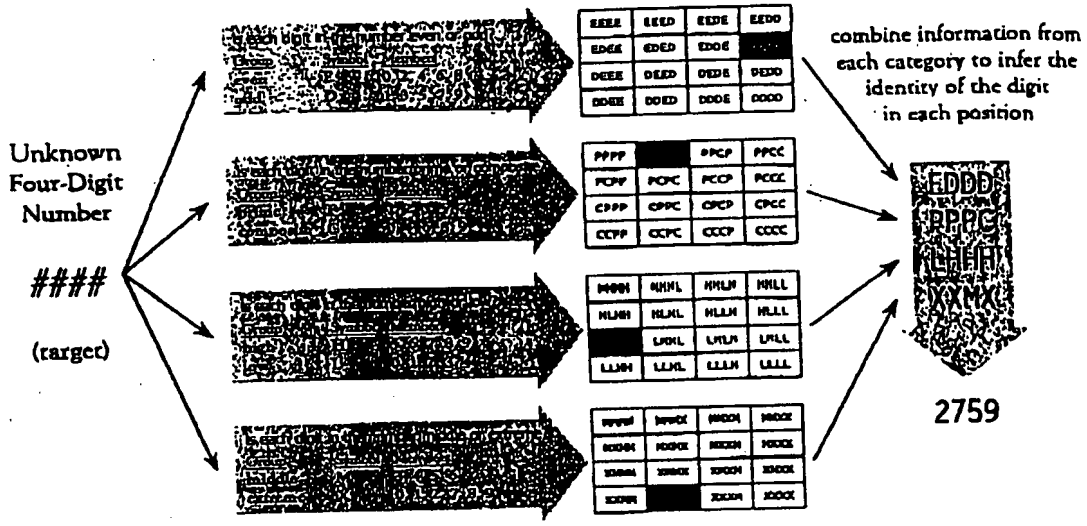
Unknown
Four-Digit
Number
####
(target)
combine information from each category to infer the identity of the digit in each position
2759

Figure 1

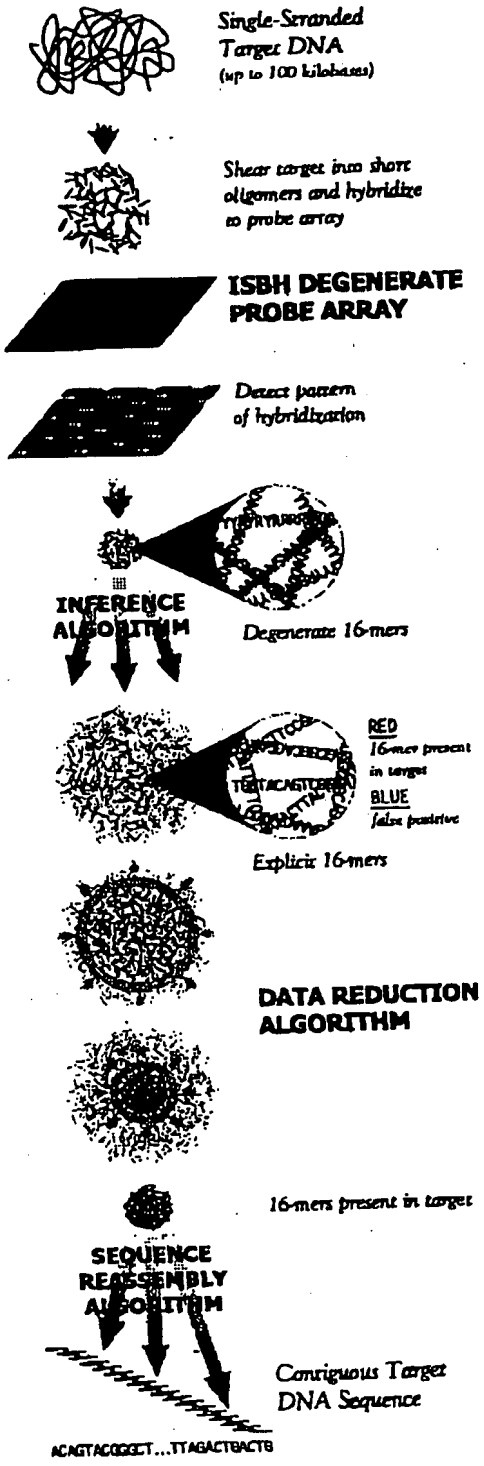
Single-Stranded
Target DNA
(up to 100 kilobases)
Shear target into short
oligomers and hybridize
to probe array
ISBH DEGENERATE
PROBE ARRAY
Detect pattern
of hybridization
INFERENCE
ALGORITHM
Degenerate 16-mers
RED
16-mer present
in target
BLUE
false positive
Explicit 16-mers
DATA REDUCTION
ALGORITHM
16-mers present in target
SEQUENCE
REASSEMBLY
ALGORITHM
Contiguous Target
DNA Sequence
ACAGTACGGGCT...TTAGACTGACTG

Figure 2

Degenerate Probe Group Pattern	*Example Probe*
$[A/B]_{16}$	ABAABBABABABABBA
$[C/D]_{16}$	DCDCDDCCCDCDCCDC
$[G/H]_{16}$	GHHHGGGHGHHGHHHG
$[T/V]_{16}$	TVTTTTVTVVVTTVTT
$[R/Y]_{16}$	RYRRYYRRYYRRYYRR
$[M/K]_{16}$	MKKKMMKMMMMKKMKM
$[W/S]_{16}$	SSWSWSSWWSWWSWSW
$[R/Y]_8[M/K]_8$	RRRRRRRRMMMMMMMM
$[M/K]_8[R/Y]_8$	MMMMMMMMRRRRRRRR
$[M/K]_8[W/S]_8$	MMMMMMMMWWWWWWWW
$[W/S]_8[M/K]_8$	WWWWWWWWMMMMMMMM
$[R/Y]_8[W/S]_8$	RRRRRRRRWWWWWWWW
$[W/S]_8[R/Y]_8$	WWWWWWWWRRRRRRRR
$[M/K]_4[R/Y]_8[M/K]_4$	MMMMRRRRRRRRMMMM
$[W/S]_4[R/Y]_8[W/S]_4$	WWWWRRRRRRRRWWWW
$[W/S]_4[M/K]_8[W/S]_4$	WWWWMMMMMMMMWWWW
$[R/Y]_4[M/K]_8[R/Y]_4$	RRRRMMMMMMMMRRRR
$[R/Y]_4[W/S]_4[R/Y]_4$	RRRRWWWWWWWWRRRR
$[M/K]_4[W/S]_8[M/K]_4$	MMMMWWWWWWWWMMMM
$([R/Y][M/K])_8$	RMRMRMRMRMRMRMRM
$([W/S][M/K])_8$	WMWMWMWMWMWMWMWM
$([R/Y][W/S])_8$	RWRWRWRWRWRWRWRW
$([R/Y]_4[M/K]_4)_2$	RRRRMMMMRRRRMMMM
$([R/Y]_4[W/S]_4)_2$	RRRRWWWWRRRRWWWW
$([W/S]_4[M/K]_4)_2$	WWWWMMMMWWWWMMMM

R = A or G M = A or C W = A or T B = C or G or T
Y = C or T K = G or T S = C or G D = A or G or T
H = A or C or T
V = A or C or G

Figure 3

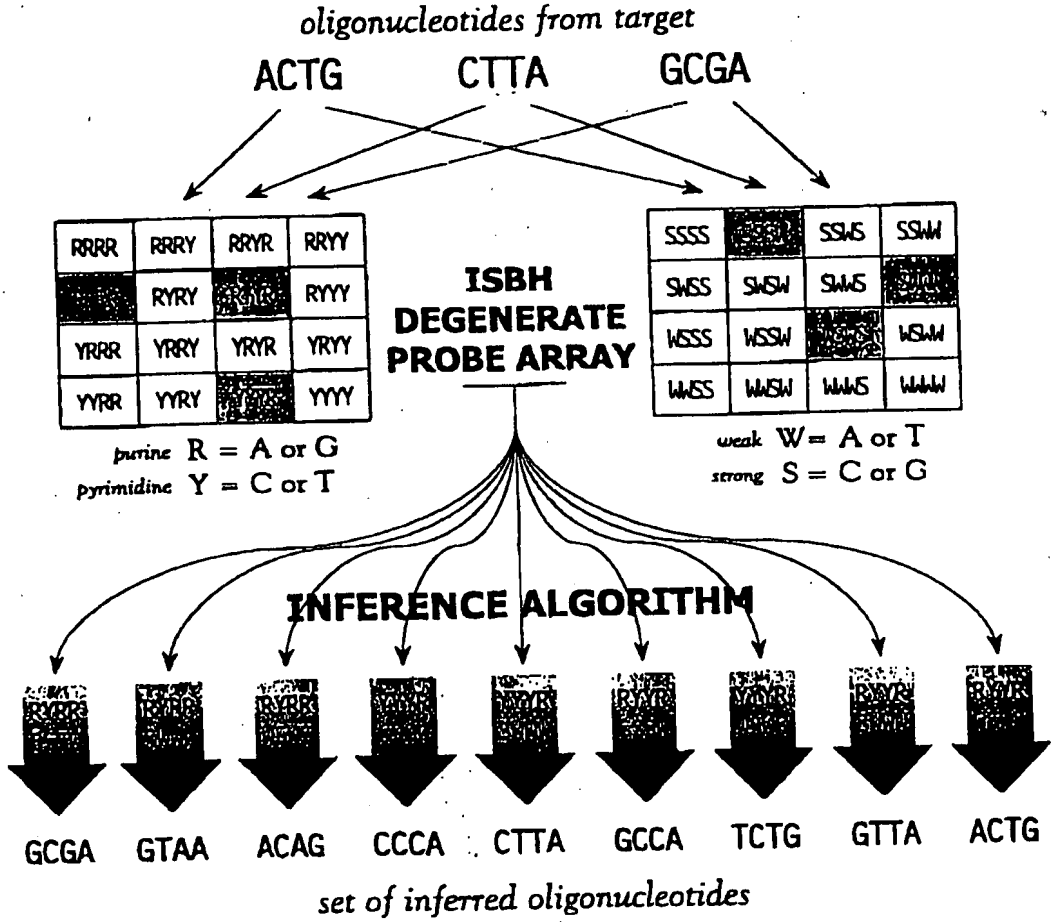
oligonucleotides from target
ACTG
CTTA
GCGA
RRRR
RRRY
RRYR
RRYY
RYRY
RYYY
YRRR
YRRY
YRYR
YRYY
YYRR
YYRY
YYYY
purine R = A or G
pyrimidine Y = C or T
ISBH
DEGENERATE
PROBE ARRAY
SSSS
SSSW
SSWS
SSWW
SWSS
SWSW
SWWS
WSSS
WSSW
WSWW
WWSS
WWSW
WWWS
WWWW
weak W = A or T
strong S = C or G
INFERENCE ALGORITHM
GCGA
GTAA
ACAG
CCCA
CTTA
GCCA
TCTG
GTTA
ACTG
set of inferred oligonucleotides

Figure 4

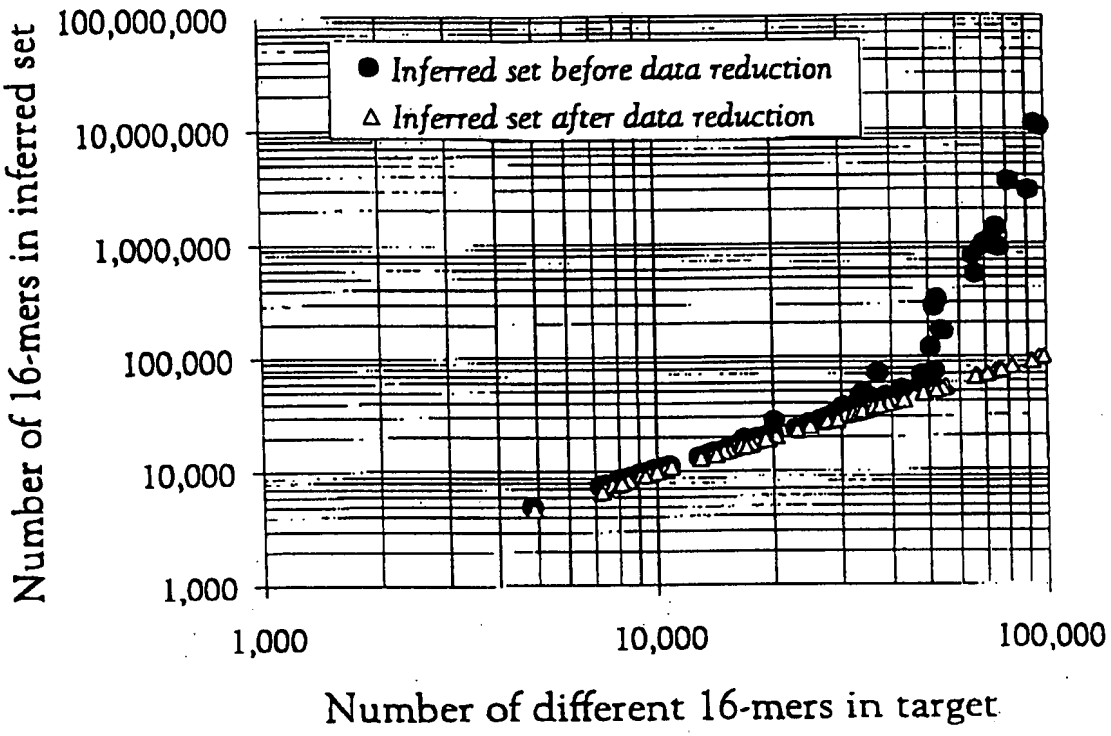
Number of 16-mers in inferred set
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
Inferred set before data reduction
Inferred set after data reduction
1,000
10,000
100,000
Number of different 16-mers in target

Figure 5

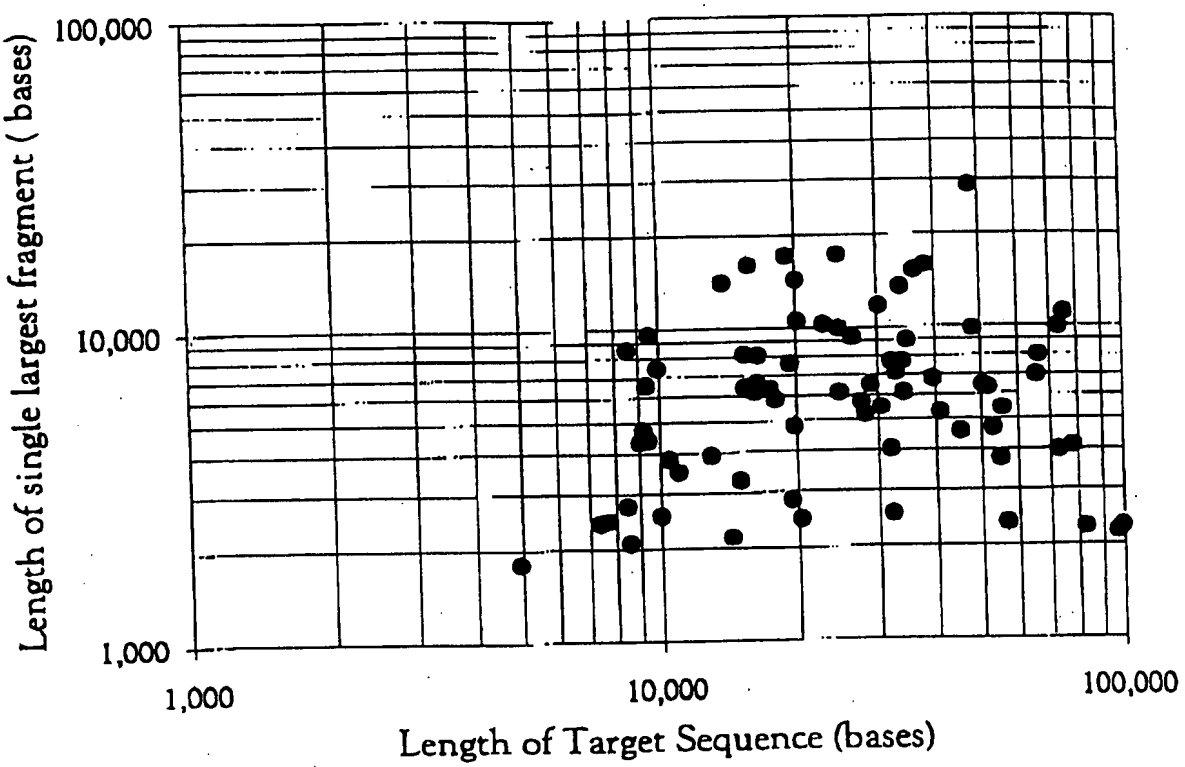
100,000
10,000
1,000
Length of single largest fragment (bases)
1,000
10,000
100,000
Length of Target Sequence (bases)

Figure 6

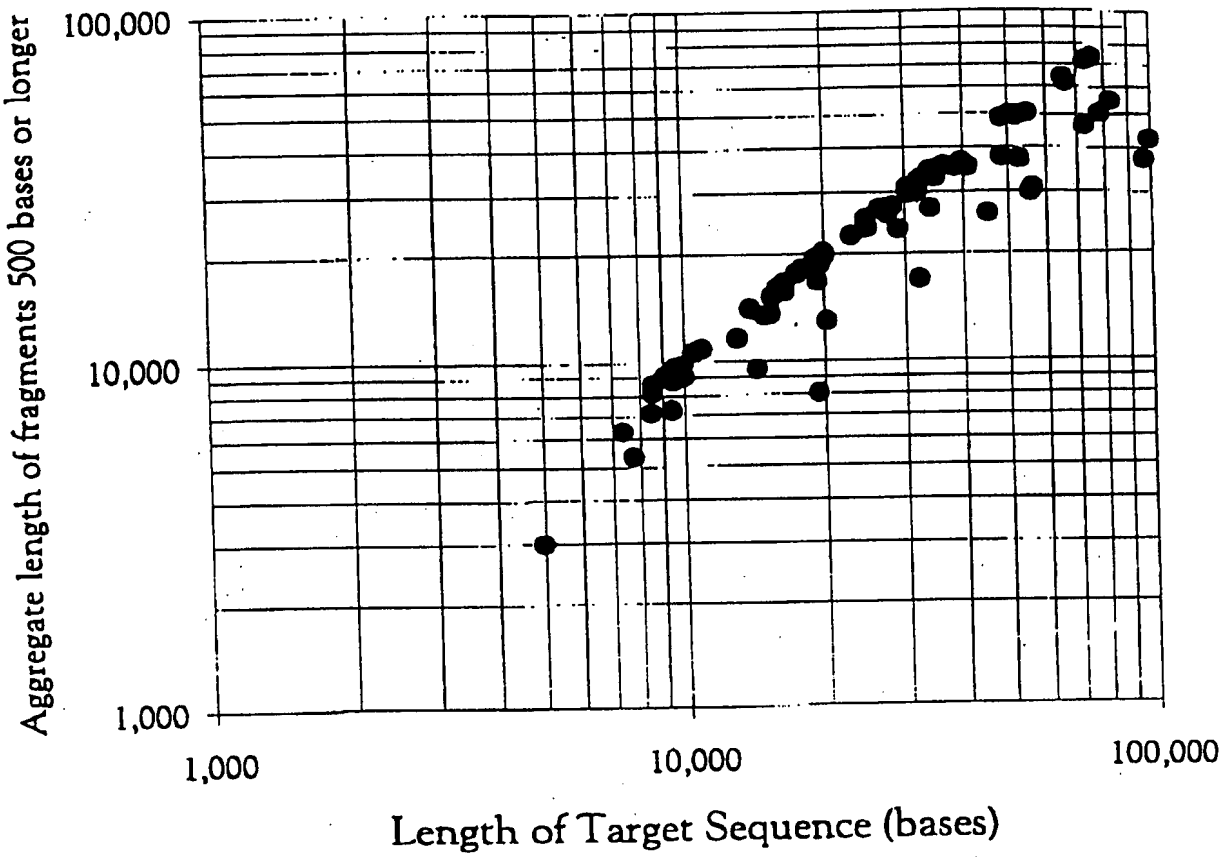
Aggregate length of fragments 500 bases or longer
100,000
10,000
1,000
1,000
10,000
100,000
Length of Target Sequence (bases)

Figure 7

Locus and Definition	Length	No. of different 16-mers in target	Fraction of Target 16-mers that are repeated	No. of 16-mers in Inferred Set	No. of 16-mers in Inferred Set after data reduction	Fraction of Inferred Set that are false positives	No. of Fragments 500 bases or longer	Median Length of Fragments 500 bases or longer	Aggregate Length of Fragments 500 bases or longer	No. of Fragments 100 bases or longer	Median Length of Fragments 100 bases or longer	Aggregate Length of Fragments 100 bases or longer	Target Sequence Coverage by Fragments at least 100 bases in length (%)	Lengths of Ten Largest Fragments
[HUMNEUROF] Human oligodendrocyte myelin glycoprotein (OMG) exons 1-2; neurofibromatosis 1 (NF1) exons 28-49; ecotropic viral integration site 2B (EVI2B) exons 1-2; ecotropic viral integration site 2A (EVI2A) exons 1-2; adenylate kinase (AK3) exons 1-2.	100849	98110	0.827	10107758	98294	0.998	43	625	23624	272	279	79656	78.58	1454 1449 1134 1123 976 960 854 837 779 768
[HSMHC3W36A] Human HLA class III region containing cAMP response element binding protein-related protein (CREB-RP) and tenascin X (tenascin-X) genes, complete cds, complete sequence.	100263	91860	0.094	2882652	68792	0.958	47	760	42076	155	358	69380	69.17	2291 1558 1514 1475 1346 1346 1281 1261 1231 1217
[F19K23] Sequence of BAC F19K23 from Arabidopsis thaliana chromosome 1, complete sequence.	97451	95804	0.027	10787002	96037	0.991	65	705	36743	243	207	76951	78.96	2198 1863 1694 1474 1367 1230 1214 1036 997 952
[HSAC000389] Genomic sequence from Human 17, complete sequence.	83024	81724	0.035	3436511	81277	0.976	57	844	54304	141	364	73331	68.33	2265 1904 1694 1559 1556 1542 1470 1463 1439 1431
[AC000400] Genomic sequence from Mouse 19, complete sequence.	74748	76621	0.027	883615	75463	0.813	44	942	50045	130	376	64161	86.56	4078 2859 2571 2312 2265 1907 1873 1698 1647 1523
[ECOHU47] 47 to 48 centisome region of E.coli K12 BHB2600.	75888	75810	0.001	1333669	75675	0.943	28	1860	73344	39	1252	75367	99.31	10886 8530 5587 5466 4471 3904 3739 3321 3211 3052
[DMAC001646] Drosophila melanogaster (P1 DS01845 (D07)) DNA sequence, complete sequence.	73727	73667	0.001	1062018	73543	0.931	28	1501	71246	36	1183	73383	99.53	9772 7742 6986 6481 4277 3549 2838 2717 2604 2319
[HUMHBB] Human beta globin region on chromosome 11.	73308	70570	0.037	962800	69925	0.927	41	660	46217	121	288	63562	85.34	3947 3297 2501 2141 2132 1822 1638 1588 1543 1417
[DMAC001653] Drosophila melanogaster (P1 DS07874 (D14)) DNA sequence, complete sequence.	68990	68907	0.001	522150	66708	0.872	39	1173	61559	56	866	66282	96.94	7947 3836 3490 3304 3062 2166 2161 2141 2060 1877
[SCE9537] Saccharomyces cerevisiae chromosome V cosmids 9537, 9581, 9495, 9867, and lambda clone 5898.	66030	65870	0.092	756772	65716	0.913	27	1984	63548	36	1776	65544	99.26	6849 4888 4289 4144 3505 3463 3335 2988 2858 2384
[HUMHPRTB] Human hypoxanthine phosphoribosyltransferase (HPRT) gene, complete cds.	56737	54261	0.043	166091	52680	0.673	33	744	30271	68	363	43560	76.77	2341 2322 1590 1532 1313 1248 1233 1217 1103 1068
[HSAC002117] Genomic sequence from Human 17, complete sequence.	55886	53117	0.067	71254	50119	0.269	22	1020	29677	59	398	39153	70.06	5374 3370 2826 1770 1536 1452 1343 1301 1277 1089
[DMAC001655] Drosophila melanogaster (P1 DS07700 (D17)) DNA sequence, complete sequence.	55188	55080	0.002	163258	54873	0.663	33	1233	49914	48	717	54746	99.2	3722 3180 3058 2874 2803 2653 2625 2276 1840 1827
[EAU67194] Enterobacter aerogenes plasmid R751, complete genome.	53331	53215	0.002	215560	52912	0.831	33	340	36277	93	368	51588	96.71	6684 2397 2087 1752 1672 1581 1566 1355 1235 1181
[MLCGA] Mycobacteriophage L5 complete genome.	52237	52260	0.000	768668	52119	0.808	11	1553	48972	34	1156	52068	99.54	6268 6182 4612 3827 3586 3423 2909 2483 1828 1619
[SCE9669] Saccharomyces cerevisiae chromosome V cosmids 9669, 8334, 8199, and lambda clone 1160.	50949	50811	0.000	116818	50688	0.565	24	1838	49840	28	1517	50608	99.32	6368 5487 3999 3616 3377 2287 2208 2137 2048 2020
[HSAF002996] Homo sapiens cosmid, complete sequence.	48874	48231	0.007	66197	47963	0.771	17	2160	37433	92	354	46468	95.64	9748 3767 3100 2487 2471 2433 2420 2310 2160 1578
[LAMCG] Bacteriophage lambda, complete genome.	48502	48447	0.000	55382	48423	0.124	3	10460	48426	3	10460	48426	99.84	28650 10460 9458

Figure 8 (1 of 4)

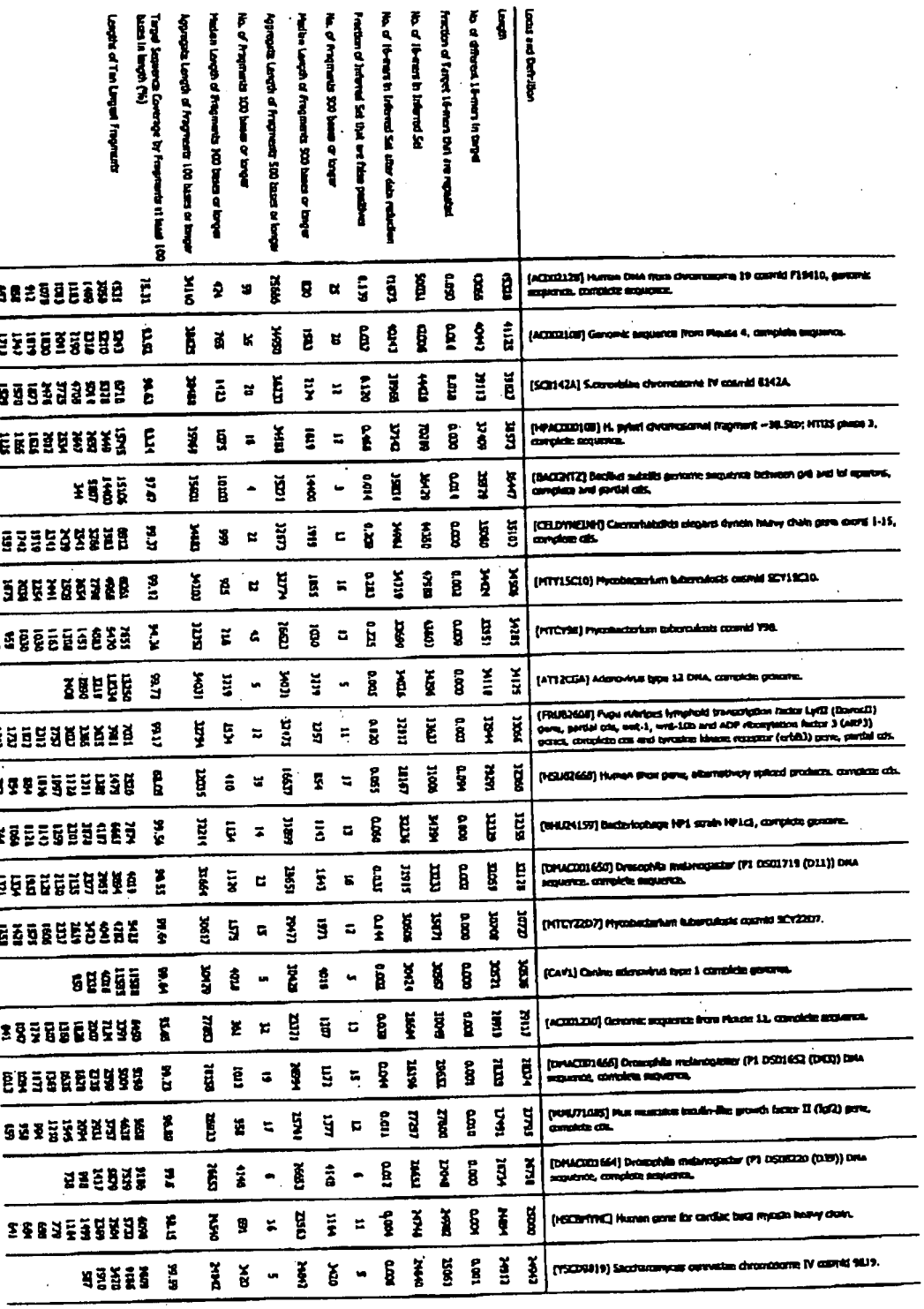

Figure 8 (2 of 4)

Locus and Definition	Length	No. of different 16-mers in target	Fraction of Target 16-mers that are repeated	No. of 16-mers in Inferred Set	No. of 16-mers in Inferred Set after data reduction	Fraction of Inferred Set that are false positives	No. of Fragments 500 bases or longer	Median Length of Fragments 500 bases or longer	Aggregate Length of Fragments 500 bases or longer	No. of Fragments 100 bases or longer	Median Length of Fragments 100 bases or longer	Aggregate Length of Fragments 100 bases or longer	Target Sequence Coverage by Fragments at least 100 bases in length (%)	Lengths of Ten Largest Fragments
[PPU24215] Pseudomonas putida p-cymene catabolism (cym) and p-cumate catabolism (cmt) operons and enol-coenzyme A hydratase gene, complete cds.	24937	24914	0.000	24925	24822	0.000	3	5398	23856	5	1634	24819	99.52	16876 5398 1634 445 436
[FRU63920] Fugu rubripes platelet-derived growth factor receptor beta (PDGFRbeta) and macrophage colony-stimulating factor receptor (CSF1R) genes, complete cds.	23156	23021	0.007	23054	22917	0.001	3	8042	22248	5	4123	22725	97.85	10117 8041 4123 352 111
[CELTPA1A] Caenorhabditis elegans gene for TPA-1A; TPA-1B, complete cds.	20381	20234	0.006	26344	19985	0.231	15	753	12827	36	415	18422	90.38	2446 1052 1028 990 901 810 764 753 732 715
[VMVGEN] Variola minor virus complete genome.	20259	20200	0.002	20412	20107	0.010	5	3384	19744	6	3276	20131	99.28	18399 4171 3381 1171 851 381
[HHSPHOS] Human herpesvirus alkaline exonuclease (CHSR), virion protein (CHNA), integral membrane protein (CHSL), origin-binding protein (CHSR), CB1R, CB2L, virion protein (CB3L), helicase (CB4L), CB5R, CB6L, phosphotransferase (CB7R, CB8R)...	20155	20131	0.000	20149	20053	0.001	2	10038	20045	2	10038	20045	99.44	14083 5991
[HSIGCHUDE] Human immunoglobulin C(mu) and C(delta) heavy chain genes (constant regions).	19795	19374	0.021	19452	18272	0.004	7	3409	18358	10	1378	19089	96.42	4773 4404 3083 2409 2013 1145 620 351 213 169
[DMACDEUH7] Drosophila melanogaster (P1 DS09219 (D101)) DNA sequence, complete sequence.	19541	19005	0.025	19222	18488	0.030	8	933	16536	15	550	18533	94.83	7589 3914 3100 945 919 681 760 550 491 461
[DMU37541] Drosophila melanogaster mitochondrial genome, complete sequence.	19517	16990	0.138	18636	16571	0.688	8	731	7988	33	339	14138	72.43	2208 1306 822 803 660 597 567 543 491 474
[CTU16304] Citrus tristeza virus complete genome.	19296	19281	0.000	19281	19185	0.000	2	9607	19186	2	9607	19186	99.42	16932 2282
[CEU26736] Caenorhabditis elegans dystrophin-like protein MEC-2 (mec-2) gene, complete cds.	18124	18108	0.003	18512	17997	0.021	7	1731	18003	7	1731	18003	99.32	5418 4228 3229 1731 1426 833 828
[XELMTCG] Xenopus laevis mitochondrial DNA, complete genome.	17553	17507	0.002	17519	17437	0.001	3	6246	17379	3	6246	17379	99	6326 6246 4849
[HUMMTA] Human mitochondrial DNA, complete sequence.	16589	16544	0.000	16549	16468	0.000	3	5363	16472	3	5363	16472	99.44	8078 5363 3073
[CELUNC52X] Caenorhabditis elegans basement membrane proteoglycan (unc-52) homologous gene, exons 1-25, complete cds.	16529	16508	0.000	16567	16395	0.004	8	1124	15483	13	915	16402	99.22	6894 2068 2066 1211 1036 917 915 783 299 248
[MUSMTCG] Mouse mitochondrion, complete genome.	16295	16279	0.000	16294	16190	0.001	7	2242	16196	7	2243	16196	99.38	6148 2082 2023 2243 910 566 521
[BSU51115] Bacillus subtilis CotA (cotA), GabP (gabP), YeaB (yeaB), YeaC (yeaC), YebA (yebA), GMP synthetase (guaA) genes, complete cds, and AIR carboxylase I (purE) gene, partial cds.	15897	15882	0.000	15883	15756	0.000	1	15811	15797	1	15811	15797	99.38	15811
[U00094] Mycoplasma genitalium tufB, rpsA, rrnA 5S, ByC genes from bases 174532 to 190046 (section 16 of 56) of the complete genome.	15495	15479	0.000	15554	15389	0.005	4	2979	15002	5	1583	15260	99.33	8182 4375 1583 908 405
[SASSV1XX] Sulfolobus virus 1 complete genome (provirus).	15465	15382	0.004	15383	15251	0.000	4	3067	13447	11	348	15004	97.01	6364 4940 1198 584 400 342 281 253 143 139
[HUMATPGG] Human gastric (H+ + K+)-ATPase gene, complete cds.	15067	14900	0.008	14951	14780	0.001	8	1371	13199	13	681	14139	93.95	3234 2951 2030 1819 1123 1018 688 654 426 198
[CELTMY1A] Caenorhabditis elegans tmy-1 gene for tropomyosin isoforms, complete sequence.	14408	14228	0.011	14263	14015	0.003	11	619	9395	27	428	13405	90	2148 1430 1179 789 622 619 581 575 574 520
[RICPPA] Rice endogenous double-stranded RNA encoding polyprotein (containing putative helicase and putative RNA-dependent RNA polymerase domains), complete cds.	13952	13937	0.000	13943	13855	0.000	1	13870	13856	1	13870	13856	99.7	13870
[U75644] Lycopersicon esculentum farnesyl-protein transferase beta subunit gene, complete cds.	13066	12999	0.004	13009	12868	0.000	6	1738	11180	9	1054	12328	94.41	3862 2480 1385 1211 1054 686 359 363 131

Figure 8 (3 of 4)

Locus and Definition	Length	No. of different 16-mers in target	Fraction of Target 16-mers that are repeated	No. of 16-mers in Inferred Set	No. of 16-mers in Inferred Set after data reduction	Fraction of Inferred Set that are false positives	No. of Fragments 500 bases or longer	Median Length of Fragments 500 bases or longer	Aggregate Length of Fragments 500 bases or longer	No. of Fragments 100 bases or longer	Median Length of Fragments 100 bases or longer	Aggregate Length of Fragments 100 bases or longer	Target Sequence Coverage by Fragments at least 100 bases in length (%)	Lengths of Ten Largest Fragments
[MUSMVDPA] Mus musculus (clone lambda-17a, lambda-20b) aldose reductase (MVDP) gene, complete cds.	11026	10946	0.006	10961	10832	0.001	6	1414	10752	6	1414	10752	97.5	3466, 2281, 1567, 1263, 1101, 1058
[AF004325] Streptococcus pneumoniae capsular serotype 19B cps19A biosynthesis locus, cps19bF gene, partial cds, cps19bG, cps19bH, cps19bP, cps19bI, cps19bQ, cps19bR, cps19bJ and cps19bK genes, complete cds, and cps19bL gene, partial cds.	10549	10534	0.000	10537	10446	0.000	3	3796	10440	3	3796	10440	99.03	3798, 3796, 2887
[XELRLSPH] Xenopus laevis genes for sperm-specific nuclear basic proteins (SPH), complete cds.	10165	7270	0.284	7287	6785	0.002	0	0	0	12	168	2415	23.75	459, 404, 270, 224, 210, 173, 158, 154, 143, 142
[D88193] Brassica rapa DNA for S-receptor kinase, complete cds.	10075	10049	0.001	10073	9952	0.002	8	1334	9023	9	755	9822	97.47	2516, 1466, 1535, 1133, 755, 702, 425, 260, 156
[AF002207] Arabidopsis thaliana kinesin-like protein (ZWICHEL) gene, complete cds.	10000	9984	0.000	9991	9876	0.001	2	4815	9603	3	2222	9879	98.77	7488, 2222, 201
[ATHACCDHCAR] Arabidopsis thaliana acetyl-CoA carboxylase gene, exons 1-31, complete cds.	9581	9566	0.000	9566	9482	0.000	1	9497	9483	1	9497	9483	98.96	9497
[CLOACOX] Clostridium magnum acoR gene, complete cds; TPP-dependent acetoin dehydrogenase alpha and beta subunit (acoAB) genes, complete cds; acoX gene, dihydrolipoamide acetyltransferase (acoCL) genes, complete cds.	9500	9330	0.016	9334	9230	0.000	4	1692	8732	4	1692	8732	91.9	4356, 1907, 1477, 1048
[ZMU41000] Zea mays copia-type retroelement PREM-2 gag gene, complete cds.	9449	8150	0.136	8151	8107	0.000	1	3624	7220	6	205	7780	82.32	6543, 735, 233, 177, 104, 102
[ZMU09989] Zea mays DJL H(+)-transporting ATPase (MHA1) gene, complete cds.	9285	9222	0.005	9224	9116	0.000	4	1898	9067	4	1898	9067	97.43	4690, 2150, 1830, 637
[HIVNDK] Human immunodeficiency virus type 1 (HIV-1) proviral complete genome.	9143	9061	0.007	9063	9046	0.000	3	4264	9024	3	4264	9024	98.66	4294, 4264, 508
[XELHISUK3] X.laevis histone H1B, H2A, H2B, and H4 genes, complete cds, and histone H3 gene, 3' end, gene cluster X1h1.	8608	8484	0.013	8493	8361	0.001	7	1030	8060	8	947	8278	96.14	2073, 1505, 1435, 1030, 664, 655, 594, 231
[MMTVPROVR] Mouse mammary tumor virus proviral DNA (from Mus musculus mammary gland) for gag-protease-pol polyprotein and env protein, complete cds.	8603	6588	0.000	6588	8490	0.000	1	8505	8491	1	8505	8491	98.68	8505
[FRU63807] Fugu rubripes growth hormone (GH) gene, complete cds.	8521	7995	0.061	8000	7836	0.001	4	1630	7097	7	1179	7623	89.37	2714, 2066, 1194, 1179, 225, 172, 171
[XLU67076] Xenopus laevis zinc finger protein XFDL156 gene, complete cds.	7782	7723	0.006	7726	7609	0.000	4	1109	5287	11	356	7171	92.11	2442, 1373, 945, 683, 410, 356, 339, 329, 288, 152
[XELEF1AZ] Xenopus laevis elongation factor 1-alpha-O gene, complete cds.	7394	7348	0.004	7354	7234	0.001	4	1574	6250	7	759	6943	93.88	2399, 1935, 1213, 759, 328, 204, 203
[HSU01102] Human lung Clara cells 10 kda secretory protein (CC10) gene, uteroglobin and Alu repeat sequences, complete cds.	4995	4861	0.020	4884	4726	0.001	3	689	2983	9	194	4034	80.33	1782, 689, 552, 320, 194, 182, 148, 140, 130

Figure 8 (4 of 4)